Why we learn less from observing outgroups

Kang, Pyungwon ; Burke, Christopher John ; Tobler, Philippe N ; Hein, Grit

Abstract: Humans are less likely to learn from individuals belonging to a different group (outgroup) than from individuals of their own group (ingroup), yet the source of this societally relevant deficit has remained unclear. Here we used neuroimaging and computational modeling to investigate how people learn from observing the actions and outcomes of ingroup and outgroup demonstrators. Politically left-wing male and female participants performed worse when observing computer-simulated actions they believed were from a right-wing outgroup member compared with those from a left-wing ingroup member. A control experiment in which participants observed choices from a nonhuman agent confirmed that this performance difference reflected an outgroup deficit, rather than an ingroup gain. Accounting for the outgroup deficit, a computational model showed that participants relied less on information from outgroup actions compared with ingroup actions, while learning from outgroup outcomes was not impaired. At the neural level, the differences in observational ingroup versus outgroup learning were reflected in lateral prefrontal activity. The stronger the activity in this region, the more strongly participants weighed ingroup compared with outgroup learning signals (action prediction errors), which formally captured deficits in outgroup learning. Together, our work provides a computational and neural account of why people learn less from observing outgroups.

DOI: https://doi.org/10.1523/jneurosci.0926-20.2020

JNeurosci
THE JOURNAL OF NEUROSCIENCE

Research Articles: Behavioral/Cognitive

Why We Learn Less from Observing Outgroups

https://doi.org/10.1523/JNEUROSCI.0926-20.2020

Cite as: J. Neurosci 2020; 10.1523/JNEUROSCI.0926-20.2020

Received: 14 April 2020
Revised: 2 November 2020
Accepted: 4 November 2020

Why we learn less from observing outgroups

Running title: Outgroup deficit in observational learning

Pyungwon Kang[1], Christopher J. Burke[1], Philippe N. Tobler[1]*, & Grit Hein[2]*

[1] Department of Economics and Laboratory for Social and Neural Systems Research, University of Zurich

[2] Translational Social Neuroscience Unit, Department of Psychiatry, Psychosomatic and Psychotherapy, University of Würzburg

* shared senior authorship

Correspondence to: pyungwon.kang@gmail.com, Blumlisalpstrasse 10, 8006, Zurich, Switzerland

Number of pages: 34

Number of figures: 6

Number of tables: 2

Number of words for abstract: 203

Number of words for introduction: 642

Number of words for discussion: 1395

Conflict of interest statement: The authors declare no competing financial interests.

Acknowledgements: This work was supported by the Swiss NSF (IZKSZ3_162109 and 100019_176016 to PNT). We also acknowledge funding by the German Research Foundation (HE 4566/5-1; HE 4566/3-1 to GH).

Abstract

Humans are less likely to learn from individuals belonging to a different group (outgroup) than from individuals of their own group (ingroup), yet the source of this societally relevant deficit has remained unclear. Here we used neuroimaging and computational modeling to investigate how people learn from observing the actions and outcomes of ingroup and outgroup demonstrators. Politically left-wing male and female participants performed worse when observing computer-simulated actions they believed were from a right-wing outgroup member compared to those from a left-wing ingroup member. A control experiment in which participants observed choices from a non-human agent confirmed that this performance difference reflected an outgroup deficit, rather than an ingroup gain. Accounting for the outgroup deficit, a computational model showed that participants relied less on information from outgroup actions compared to ingroup actions, while learning from outgroup outcomes was not impaired. At the neural level, the differences in observational ingroup vs outgroup learning were reflected in lateral prefrontal activity. The stronger the activity in this region, the more strongly participants weighed ingroup compared to outgroup learning signals (action prediction errors), which formally captured deficits in outgroup learning. Together, our work provides a computational and neural account of why people learn less from observing outgroups.

Significance statement

Learning from observing others is an efficient way to acquire knowledge. In our globalized world, “the others” often are people from a different social group (outgroup). There is evidence that people learn less from observing outgroup individuals compared to individuals from their own group (ingroup). However, the source of this outgroup deficit in observational learning remained unknown, which limits our chances to improve intergroup learning. Our results showed that participants rely less on observed outgroup actions compared to ingroup actions, while learning from outgroup outcomes is not impaired. On the neural level, this outgroup deficit was reflected in activation of the inferior frontal gyrus. These findings imply that intergroup learning should rely on observing outcomes, rather than actions.

Introduction

For many different species, including humans, learning from perceiving the actions and outcomes of others (i.e., observational learning) is an efficient way to acquire knowledge and skills. There is evidence that observational learning is modulated by important social factors such as group membership (Buttelmann et al., 2013; Golkar et al., 2015; Howard et al., 2015). For example, learning appears to be facilitated if participants observe a person from their own social group (ingroup) compared to a person from a different social group (outgroup) (Golkar et al., 2015; Golkar and Olsson, 2017). This ingroup bias in observational learning was even found in infants and children being more likely to imitate the novel actions of a demonstrator who speaks their language, compared with a person speaking a different language (Buttelmann et al., 2013; Howard et al., 2015).

Previous neuroscience studies have investigated observational learning irrespective of group membership (Burke et al., 2010; Suzuki et al., 2012; Charpentier et al., 2020; Kumaran et al., 2015) and reported learning from observed outcomes and from observed actions. Learning from others' outcomes and the resulting outcome prediction errors is associated with activation in medial prefrontal cortex (MPFC; Burke et al., 2010; Suzuki et al., 2012; Kumaran et al., 2015). Learning from others' actions and the resulting action prediction errors is related to activation in dorsolateral prefrontal cortex/inferior frontal gyrus (DLPFC/IFG; (Burke et al., 2010; Suzuki et al., 2012; Charpentier et al., 2020).

The effect of group membership has mainly been investigated with regard to action observation or imitation (Losin et al., 2012; Losin and Woo, 2015), revealing ingroup vs outgroup differences in brain regions associated with mentalizing (Losin and Woo, 2015), and parts of the 'mirror neuron system' such as the inferior frontal, motor and parietal cortex (Losin

et al., 2012). However, it remained unclear how important social factors, such as group membership, shape observational learning mechanisms and the underlying neural circuitries.

In our study, we investigated how group membership affects the neural circuitries of observational learning. To do so, we combined a well-established observational learning paradigm (Burke et al., 2010) with social group manipulation, computational modeling and fMRI. In more detail, participants inside the fMRI scanner observed only choices (i.e., actions) or choices and outcomes of an ingroup and an outgroup demonstrator, and could use these different pieces of information to optimize their own choice.

Based on previous behavioral evidence of outgroup deficits in social learning (Buttelmann et al., 2013; Golkar and Olsson, 2017), we hypothesized that participants choose the "correct" (i.e., more rewarding) option less frequently after observing outgroup choices compared to ingroup choices, reflecting an outgroup deficit in observational learning. Given that individuals learn from observing A) the outcomes, and B) the actions of others (Burke et al., 2010; Suzuki et al., 2012), we derived three different hypotheses regarding the mechanisms that might underlie the potential deficit in outgroup learning. According to a first hypothesis, the outgroup deficit in observational learning may arise because participants rely more on observed ingroup compared to outgroup outcomes. In computational modeling, this should be reflected by a stronger weight for ingroup compared to outgroup outcome prediction errors, associated with neural activation of the MPFC (Burke et al., 2010; Suzuki et al., 2012; Kumaran et al., 2015). Alternatively, the outgroup deficit in observational learning may occur because participants rely more on observed ingroup compared to outgroup actions. In this case, our computational modeling results should reveal a stronger weight for ingroup compared to outgroup action prediction errors, related to increased activation in DLPFC/IFG (Burke et al., 2010; Suzuki et al.,

2012; Charpentier et al., 2020). Finally, it is possible that the outgroup deficit in observational learning arises because participants rely more on observing ingroup outcomes and actions, reflected by a stronger weight for ingroup compared to outgroup outcome and action prediction errors, paralleled by increased activation in MPFC and DLPFC/IFG.

Methods

Participants

fMRI study. Thirty-two participants (19 female, mean age: 22.51±0.54 years) were recruited from the University of Zurich and a local community in Zurich. Participants were all right-handed, had normal/corrected-to-normal vision and did not have a history of psychological or neurological disorder. Because our group manipulation was based on political attitude (see below), we invited Swiss participants who perceived themselves as politically active (i.e., interested in current political debates in Switzerland) with primarily left-wing attitudes. Three participants had to be excluded because they showed right-wing attitudes during the group manipulation check (see below). Thus, we analyzed data of 29 participants in the imaging experiment.

Behavioral control study. For a behavioral control experiment outside the fMRI scanner, we recruited another sample of 33 participants (female: 18, mean age: 23.25±3.34). The participants were recruited from the same participant pool and they matched the participants of the fMRI study in terms of age, education level, political attitude, and nationality, all ps > 0.171. All participants received a fixed monetary compensation for their participation and additional incentives according to their performance. The study was approved by the ethics committee of the Canton of Zurich.

144

145 **Experimental design and statistical analysis**

146 fMRI study. *Prescanning procedure for group induction.* Before the main fMRI experiment,
147 participants provided their political views by rating current political issues in Switzerland, and
148 observed the ostensible ratings of two other individuals. One of these individuals displayed
149 similar ratings as the participant, indicating a left-wing attitude. The ratings of the other
150 individual indicated attitudes opposite to those of the participants (i.e., right-wing attitudes).

151 The prescanning procedure consisted of six trials. In each of these trials the participants
152 were presented with a political initiative dividing the left- and right-wing parties (e.g., an
153 initiative to raise inheritance tax). They were asked to indicate their opinion by moving an
154 abstract symbol on a visual analogous scale (ranging from "strongly disagree" to "strongly
155 agree"; **Figure 1A**). Next, participants were presented with two abstract symbols on the same
156 rating scale that ostensibly indicated the ratings of two different individuals (**Figure 1A)**. One of
157 these symbols (designated to become the symbol of the ingroup demonstrator) appeared in the
158 part of the rating scales that indicated agreement with left-wing initiatives and disagreement with
159 right-wing initiatives, corresponding to a political attitude similar to that of the participants. The
160 other symbol (designated to become the symbol of the outgroup demonstrator) appeared in the
161 part of the rating scales that indicated agreement with right-wing initiatives and disagreement
162 with left-wing initiatives, opposing the political attitude of the participant. The presentation order
163 of the symbols was randomized across trials. The symbols representing the ingroup and the
164 outgroup demonstrator were counterbalanced across participants, but remained constant within
165 each participant. At the end of the group induction, the ratings of participants and the two other
166 individuals (i.e., all three symbols) were again shown for each political issue to remind the

participants of the political attitudes of the other two individuals relative to their own attitude (**Figure 1A**).

Next, participants were asked to rate how close they feel to major political parties in Switzerland ranging from left-wing to right-wing, and to provide the same closeness ratings for each of the two other individuals whose ratings regarding political initiatives they had observed before. To do so, participants moved the respective symbols on a rating scale (ranging from "very close" to "not close at all"). These ratings served as manipulation check of our group manipulation because they quantified how differently the participants perceived the person associated with the ingroup and the outgroup symbol, and verified that the perceived differences resulted in social categorization (here supporters of left-wing and right-wing political parties).

Observational learning task. We used a modified version of an observational learning task established in a previous study (Burke et al., 2010). The participants were instructed to learn about the reward probability of two fractal images through observation of only choices (action_only condition) or choices and outcomes (action_outcome condition) of the two demonstrators whose political attitudes they had rated outside the scanner (**Figure 1A**). To do so, they would observe prerecorded choices that the two individuals (demonstrators) made to optimize their own outcomes in a previous experiment. Unknown to the participant, all demonstrator choices were generated by a standard reinforcement learning algorithm with a learning rate of 0.3 and a beta of 0.4 in each condition. We used this manipulation to ensure that participants observed identical choices from the ingroup and the outgroup demonstrator, action-outcome condition, $F(1,172) = 0.101$, $p = 0.751$, $\eta^2 = 0.001$ and in the action only condition, $F(1, 172) = 0.201$, $p = 0.655$, $\eta^2 = 0.001$. Moreover, estimating the actually realized betas of ingroup and outgroup demonstrators revealed no difference between conditions, $t(28) = .91$, $p = 0.36$.

Thus, participants observed comparable choices (and a comparable learning rate) in the ingroup and the outgroup conditions, and potential differences between the conditions should result from the group manipulation (i.e., the differences in the demonstrators' group membership).

Each trial of the observational learning task consisted of an observation phase (i.e. observing the demonstrator's decision) and a decision phase (i.e. making a decision for themselves; **Figure 1C**). To prevent confusion between the two phases, the screen was vertically split in two halves with one half showing the observation phase and the other half showing the decision phase. The display side of the observation and decision phase was constant within each participant, and counterbalanced across participants.

At the beginning of the observation phase, the demonstrator for the present trial was indicated by one of the two previously learned abstract symbols, presented for a jittered duration of 1-10s. Then, participants had 1.5 s to predict which option the demonstrator would choose. After response registration (0.5s) and a short interval (0.5s), the action of the demonstrator was shown for 1s. Depending on the condition, participants then observed the outcome of the demonstrator (action_outcome condition) or a pixel-matched scrambled image (action_only condition) for 1.5 s. The scrambled image was made from the outcome presentation part (0 or 10) and all pixels of the respective outcome were randomly redistributed. After a jittered interval of 1-10s, the demonstrator's symbol was replaced by the participant's symbol, followed by the two fractals (~1.5s). Within this period, the participants selected one of the two fractals. The choice was displayed for 0.5s, followed by a scrambled outcome image (1.5s), which prevented individual learning during the task. Learning was incentivized by paying out participant decisions at the end of the experiment. At the end of the experiment, participants were debriefed and filled in a questionnaire asking open questions regarding the purpose of the study and how

213 they felt during the study in the fMRI scanner. None of the participants reported suspicions about
214 the experimental setting or inferred the purpose of the study (effect of group membership on
215 observational learning) correctly.

216 The fractals were associated with different reward contingencies (70% reward vs. 30%
217 reward). We used the two conditions (action-outcome vs. action-only) to disentangle action-
218 related learning and outcome-related learning. In the action-outcome condition, participants
219 could learn from both actions and outcomes of the demonstrators, while in the action-only
220 condition they could learn only from observing the actions of the demonstrators. The
221 observational learning task consisted of three ingroup condition and three outgroup condition
222 sessions, with each session comprising one action_outcome block and one action_only block.
223 The sequence of ingroup and outgroup sessions was interleaved, and the order of sessions was
224 counterbalanced across participants.

225 *Individual learning task.* In addition to the observational learning task, participants
226 performed an individual learning task. The structure of this task was identical to the
227 observational learning task. However, now the symbol representing the demonstrator was
228 replaced by a scrambled image, and participants were asked to press a random key instead of
229 predicting the choice of the demonstrator. At the time when the demonstrator's choice was
230 revealed in the observational learning part, both options were highlighted by a frame to keep
231 action observation uninformative. Participants received the feedback of their own choice during
232 the decision phase. They performed the individual learning task in a separate block at the end of
233 the observational learning sessions. For both tasks, we used ten trials per block, which resulted in
234 150 trials in total.

Behavioral control study. In the behavioral control study, we investigated observational learning from a non-social agent i.e., from computer-generated choices. The observational learning task was identical (i.e. instruction, number of sessions, number of trials) to the observational learning task of the fMRI study described above, except that participants were told that they observed decisions generated by a computer. The comparison with a computer demonstrator enabled us to qualify the social observation effect from the fMRI study as outgroup deficit or ingroup enhancement. In addition to observational learning from the computer, participants in the control study also performed the individual learning task, for comparison with the fMRI study. Moreover, we determined participants' political attitude based on their ratings of the same political initiatives as in the fMRI study. Three participants who showed a right-wing attitude in these ratings were excluded from the analyses. To keep procedures as similar as possible to the observational learning task, we did not measure beliefs about the strategy, programming or reality of the computer generated agent.

Statistical analyses of behavioral data

All the behavioral analyses were performed using SPSS version 23.0 (IBM Corp., 2015). For most of the analyses, we used a repeated measure ANOVA or a paired-t-test because of the within-subjects design of the study except the κ analysis. When analyzing the κ of ingroup and outgroup, Wilcoxon sign-ranked tests were used due to the nonparametric nature of the data.

Computational modeling

We fitted reinforcement learning models to capture observational learning from actions and outcomes. Computational modeling was conducted using R (R Core Team, 2013) and a package

bbmle. For outcome learning, we assumed that participants track the demonstrator's internal learning process by simulating outcome prediction errors (sOPE, Eq.1) experienced by the demonstrator, (Suzuki et al., 2012). These prediction errors correspond to the difference of what the demonstrator received and the simulated value of the choice made by the demonstrator:

$$Eq\ 1.\ sOPE = R^{other}_{t^{other\_outcome}} - Qa^{other}_{t-1^{other\_outcome}}$$

For action learning, we modeled observed action prediction errors (APE, Eq.2) which relate actually observed choice (1, 0) to the learned probability of observing that choice (Burke et al., 2010).

$$Eq\ 2.\ APE = A^{other}_{t} - Pa^{Subject}_{t-1^{other\_action}}$$

To model how strongly participants rely on both of these prediction errors to influence their own choices, we use weights that update the value Q of the chosen option, similar to an individual learning rate. Specifically, ϖ denotes the degree to which particpants incorporate sOPEs to update Q. The larger the ϖ, the more heavily participants weighted the sOPEs. Conversely, the action learning weight κ denotes the degree to which participants integrate the APE into their own choice. We used Equations 3 and 4 for outcome and action learning and applied a softmax function with a perseverance parameter (range [0 3]) (Wunderlich et al., 2012) to convert value into action. Both ϖ and κ had a range of [-1 1], allowing for the possibility of reverse learning from the outgroup.

$$Eq\ 3.\ Qa^{subject}_{t^{other\_outcome}} = Qa^{subject}_{t-1^{other\_outcome}} + \varpi * \left(R^{other}_{t^{other\_outcome}} - Qa^{other}_{t-1^{other\_outcome}}\right)$$

$$Eq\ 4.\ Pa^{Subject}_{t^{other\_action}} = Pa^{Subject}_{t-1^{other\_action}} + \kappa * (A^{other}_{t} - Pa^{Subject}_{t-1^{other\_action}})$$

We assessed observational learning from the ingroup and outgroup by testing models with separate or common ϖ and κ for the ingroup and outgroup conditions, resulting in four models in total. All the models were fitted at the individual level. For model comparison, we calculated the summed maximum likelihood for all conditions and trials for each participant and used the Akaike information criterion (AIC) to determine the best model.

In addition to the decision for themselves, we modeled the predictions participants made regarding the choices of the demonstrators, using the most recent sOPE and APE (Eq 5. and Eq 6.) Again, we compared four models with separate or common ϖ and κ for the ingroup and outgroup conditions to test whether group membership differentially affects prediction learning. However, we then entered the output of these models into the softmax function to generate predictions of the decisions of the demonstrator rather than decisions for themselves.

$$Eq\ 5.\ Qa^{Subject}_{t+1^{other\_outcome}} = Qa^{Subject}_{t^{other\_outcome}} + \varpi * \left(R^{other}_{t^{other\_outcome}} - Qa^{other}_{t^{other\_outcome}}\right)$$

$$Eq\ 6.\ Pa^{Subject}_{t+1^{other\_action}} = Pa^{Subject}_{t^{other\_action}} + \kappa * \left(A^{other}_{t} - Pa^{Subject}_{t^{other\_action}}\right)$$

As an alternative model family, we considered the possibility that action prediction errors contribute to the learning process in the action-outcome condition as both action and outcome information are observable in this condition. In order to test whether adding κ-weighted action prediction errors to the action_outcome condition improves model fit, we examined the model family described in Equation 7. The weight parameter captures the relative use of sOPE and APE for updating Q. We again estimated four models varying whether ingroup and outgroup parameters were common or separate. The average model fit was worse than for the models without κ in the action_outcome condition (mean AIC: 17.58 vs. 18.56). Moreover, the best-fitting model of the alternative family explained the data less well than the best-fitting model

without κ in the action_outcome condition. We therefore used the models without κ in the action_outcome condition for further analysis.

$$Eq\ 7.\ Qa^{Subject}_{t\,other\_outcome} = Qa^{Subject}_{t-1\,other\_outcome} + \text{weight} * \varpi * \left(R^{other}_{t\,other\_outcome} - Qa^{other}_{t-1\,other\_outcome}\right) + (1 - \text{weight}) * \kappa * \left(A^{other}_{t} - Pa^{Subject}_{t-1\,other\_action}\right)$$

fMRI data acquisition and analyses

MRI data was acquired with a Philips Achieva 3T whole-body scanner (Philips Medical Systems, Best, The Netherlands) equipped with an 8-channel head coil. For each participant, we collected a T1-weighted whole brain structure image (number of slices: 181, voxel size: 1x1x1mm, field of view: 256x256mm). To measure neural activity we collected T2* weighted whole-brain echo planar images (number of slices: 40, repetition time: 2.36s, voxel size: 3 x 3 x 3mm, field of view: 256x 256mm, echo time: 30ms, flip angle: 90°).

All functional images were distortion corrected, segmented according to the individual T1 structural image, normalized, and smoothed with an 8mm isometric Gaussian kernel (full width at half maximum). Preprocessing and analyses were performed using SPM12 (Wellcome Trust Centre for Neuroimaging). To analyze functional activity, we applied a general linear model with the following regressors: 1) onset of the screen displaying the choice options for the demonstrator, 2) onset of the screen displaying the participant's prediction of demonstrator choice, 3) onset of the screen displaying the choice of the demonstrator, parametrically modulated by 4) the APE (see computational model), 5) onset of the screen displaying the outcome of the demonstrator, parametrically modulated by 6) the sOPE in the action-outcome condition (see computational model), 7) onset of the screen displaying the choice options for the

participant, 8) onset of the screen displaying the participant's choice, and 9) onset of the screen displaying the outcome/masked outcome for the participant. The duration of all events was set to 0. The six head motion regressors and a constant were included as regressors of no interest.

We assessed prediction error-related activity in a random effect model with one-sample t-tests for the contrast images created by the parametric modulators. In order to analyze APE-related activation independently of demonstrator group, we weighted both ingroup and outgroup action prediction error regressors with a 1 on the first level and used the resulting contrast images to perform a one-sample t-test against zero on the second level. The same analysis was performed to assess sOPE-related activation irrespective of group, using the respective first-level images from the ingroup and outgroup conditions. We also tested for ingroup vs outgroup differences in APE and sOPE- related activity at the first-level. Finally, using second-level correlation, we related the differences in behavioral weights (κ) given to ingroup versus outgroup action prediction errors to differential neural activity induced from observing ingroup versus outgroup demonstrator choices. We performed whole brain analyses ($p<0.05$, family wise cluster-level whole brain corrected with a cluster inducing voxel-level threshold of $p<0.001$).

Results

Group induction. Prior to scanning, participants rated their own closeness and the closeness of the future demonstrators in the observational learning task to left- and right-wing parties. A two-way ANOVA of demonstrator (ingroup/ outgroup) by party (left-wing/ right-wing) revealed neither a main effect of party ($F(1,28)=1.93.59$, $p=.17$, $\eta^2 = 0.064$) nor demonstrator ($F(1,28)=1.70$, $p=0.20$, $\eta^2 = 0.057$) but a significant interaction of demonstrator by party ($F(1,28)=259.59$, $p<0.001$, $\eta^2 = 0.903$). Participants rated themselves, ($t(28) = 12.64$, $p < 0.001$),

and the future ingroup demonstrator, ($t(28) = 15.31$, $p < 0.001$), as close to a left-wing party. The future outgroup demonstrator was rated as close to a right-wing party, ($t(28) = 13.70$, $p < 0.001$) (**Figure 1B**). The difference in closeness ratings to the left-wing party between the participants and the fellow left-wing supporter (ingroup demonstrator) were significantly smaller than the differences in closeness ratings between the participants and the person they perceived as right-wing supporter (outgroup demonstrator), $t(28) = 13.701$, $p < 0.001$. These results show that the participants perceived one of the demonstrators as a member of their own group (ingroup; defined by left-wing political attitude) and the other demonstrator as a member of a different social group (outgroup; defined by right-wing political attitude).

fMRI study

Behavioral results. In the decision phase of the observational learning task participants made more correct choices (i.e. selected the option associated with higher reward probability) after observing the ingroup compared to the outgroup demonstrator (**Figure. 2A**), group (ingroup, outgroup), $F(1, 28) = 7.839$, $p = 0.009$, $\eta^2 = 0.219$. This difference emerged over time, group x trial (one to ten) interaction, $F(9, 252) = 1.938$, $p = 0.047$, $\eta^2 = 0.065$. There was no significant difference between the action-only and the action-outcome condition, condition (action_only, action_outcome), $F(1,28) = 0.988$, $p = 0.329$, group x condition interaction, $F(1, 28) = 0.820$, $p = 0.373$, group x condition x trial interaction, $F(9,252) = 0.925$, $p = 0.504$. However, separate analyses for each condition revealed that the effect was mainly driven by the action-only condition, showing a significant main effect of group (ingroup, outgroup), $F(1,28) = 7.421$, $p = 0.011$, $\eta^2 = 0.210$ and a significant group x trial interaction, $F(9,252) = 2.327$, $p = 0.016$, $\eta^2 = 0.077$. In the action-outcome condition, we observed a marginally significant main effect of

group, F (1,28) = 3.323, p = 0.079, η^2 = 0.106, and no significant group x trial interaction p = 0.813. Compared to individual learning, participants learned less from the outgroup demonstrator, F(1,28) = 8.168, p = 0.008, η^2 = 0.226, but similarly well from the ingroup demonstrator, F(1,28) = 0.174, p = 0.714 (**Figure 2A**). Together, the results show that participants learned less from observing the outgroup compared to the ingroup demonstrator, indicating an outgroup deficit in observational learning, primarily when observing only the actions of others.

Next, we asked if the group difference we observed in choice was mirrored by a similar group difference at the prediction stage, i.e., when participants predicted the upcoming choice of the ingroup and outgroup demonstrator. To test this possibility, we conducted an ANOVA with individual choice predictions and individual choices as dependent variable, and group (ingroup/outgroup) and response type (choice prediction/choice) as independent variables. The results showed a significant main effect of group, F(1,28) = 7.472, p = 0.011, η^2 = 0.211, a significant main effect of response type, F(1,28) = 5.492, p =0.026, η^2 = 0.164, and a significant group x response type interaction, F(1,28) = 5.35, p =0.028, η^2 = 0.160 (**Figure 2B**). Clarifying this interaction effect, post-hoc pairwise comparisons showed that participants predicted the choices of the ingroup and the outgroup demonstrators equally well, t(28) =1.408, $M_{difference}$ = 0.027, SE = 0.019, p = 0.171, but showed significantly fewer correct choices in the outgroup compared to the ingroup condition, t(28) =2.800, $M_{difference}$ = 0.112, SE = 0.040, p = 0.009. Thus, participants learned to predict ingroup and outgroup choices similarly well, but used the learned information to a lesser degree when observing the outgroup compared to the ingroup demonstrator.

Behavioral control study

We conducted a behavioral control experiment to clarify whether the observed ingroup vs outgroup difference in observational learning reflects increased learning from the ingroup or reduced learning from the outgroup. The control experiment was identical to the main experiment, except that participants observed choices from a computer, i.e., a non-human agent. In case of increased learning from the ingroup demonstrator, the number of correct choices should be significantly higher in the ingroup demonstrator condition compared to the computer condition. Conversely, in case of decreased learning from the outgroup demonstrator, the number of correct choices should be significantly lower in the outgroup demonstrator condition compared to the computer condition. To test this issue, we performed two repeated measures ANOVAs with trials and condition (action_only/action_outcome) as within subject variables and demonstrator (ingroup/outgroup, computer) as a between subject variable. The results showed significantly fewer correct choices after observing the outgroup demonstrator, compared to the computer, demonstrator, $F(1,57) = 14.343$, $p = 0.0003$, $\eta^2 = 0.201$, mean difference between the outgroup and the computer conditions: $M_{\_combined} = 0.142$, SE=0.029; $M_{\_action\_outcome} = 0.135$; SE = 0.035, $M_{\_action\_only} = 0.205$, SE = 0.038 (see **Figure 2C**). There were no other significant effects, condition (action_only / action_outcome), $F(1,57) = 0.464$, $p = 0.499$, $\eta^2 = 0.008$, demonstrator x condition interaction, $F(1,57) = 1.66$, $p = 0.203$, $\eta^2 = 0.008$. In contrast to the difference between outgroup and computer demonstrator, there were no significant differences in the number of correct choices between the ingroup and the computer condition, demonstrator (ingroup / computer), $F(1,57) = 1.639$, $p = 0.208$, $\eta^2 = 0.028$, condition (action_only / action_outcome), $F(1,57) = 0.090$, $p = 0.765$, $\eta^2 = 0.002$, demonstrator x condition interaction, $F(1,57) = 0.174$, $p = 0.678$, $\eta^2 = 0.003$, mean difference between ingroup and computer conditions: $M_{\_combined} = 0.057$, SE = 0.029, $M_{\_action\_outcome} = 0.048$, SE = 0.031, $M_{\_action\_only} = 0.067$, SE =

412 0.036 (see **Figure 2C**). Individual learning in the fMRI study and control study were not
413 different, ($F(1,58) = 1.031$, $p = .314$, $\eta^2 = .016$). These results indicate that the observed group
414 difference in observational learning reflects an outgroup deficit, rather than enhanced learning
415 from the ingroup.
416
417 **Computational modeling**
418 We hypothesized that an outgroup deficit in observational learning might be driven by
419 differences in outcome-related learning, differences in action-related learning, or differences in
420 both learning mechanisms. To test these hypotheses, we fitted reinforcement learning models to
421 choice behavior when participants observed outcomes and actions from ingroup or outgroup
422 demonstrators. Model comparisons showed that behavior was best characterized by a model that
423 used a common learning weight (ϖ) for ingroup and outgroup outcome prediction errors (OPEs),
424 but separate learning weights (κ) for ingroup and outgroup action prediction errors (APEs;
425 **Figure 3A**). The κ-ingroup weight was larger than the κ-outgroup weight, Wilcoxon Rank-sum
426 test, $z = 3.13$, $p = .002$ (**Figure 3B**). Thus, model comparison results support the notion that the
427 differential observational learning effect is mainly due to reduced action-based learning from the
428 outgroup.
429 Using a similar approach and models, we performed model comparisons also for
430 participants' predictions of the demonstrators' choices. The best model for prediction behavior
431 was the model with common learning weight (ϖ) for ingroup and outgroup outcome prediction
432 errors (OPEs) and learning weight (κ) for ingroup and outgroup action prediction errors (**Figure**
433 **3C**). Thus, our participants learned to predict the choices of ingroup and outgroup demonstrators
434 similarly well (but used the acquired information differentially for their own choices).

435

436 **fMRI**

437 *Replication of previous results: Observational learning irrespective of group membership.* First,
438 we investigated if our neural results replicate the findings of previous studies that investigated
439 observational learning irrespective of group membership, showing activation in MPFC
440 associated with outcome prediction errors and activation in DLPFC related to action prediction
441 errors (Burke et al., 2010; Suzuki et al., 2012). To do so, we conducted two separate parametric
442 regression analyses that regressed the participants' trial-by-trial model estimates of ingroup and
443 outgroup outcome-prediction errors and action prediction errors against their neural activity
444 during the observation of outcomes or actions, respectively.

445 Group independent outcome-related learning activated a network of brain regions,
446 including the dorsomedial prefrontal cortex (DMPFC), bilateral insula, caudate and midbrain
447 (**Table 1, Figure 4A and B**). The neural response in these regions increased with decreasing
448 outcome-prediction errors, indicating that observed outcomes eliciting smaller outcome-
449 prediction errors resulted in stronger activity. At the applied threshold, there was no region
450 where neural response increased with increasing outcome-prediction errors. Notably, the
451 DMPFC findings co-localize with previously unreported findings of inverse outcome prediction
452 error coding in the study of Burke and colleagues (2010; **Figure 4A**).

453 Conversely, learning from observing ingroup and outgroup actions activated the anterior
454 IFG and parietal regions. (**Table 2, Figure 5A and B**). The IFG findings co-localize with those
455 of our previous report on action prediction error coding in that area (Burke et al., 2010, **Figure**
456 **5A**).

457

458 *Group differences in observational learning mechanisms.* Second, we investigated the brain
459 regions that are differentially involved in learning from ingroup and outgroup outcomes and
460 actions. To do so, we contrasted the neural responses related to participants' trial-by-trial model
461 estimates of ingroup outcome-prediction errors and action prediction errors with the neural
462 responses related to their trial-by-trial model estimates of outgroup outcome-prediction errors
463 and action prediction errors. The results revealed no significant differences at the applied
464 threshold, suggesting that participants activated similar neural circuitries while learning from the
465 observation of ingroup and outgroup outcomes and actions.
466
467 *Group differences in the weight assigned to observational learning.* Third, we tested our
468 assumption that the stronger weight assigned to action prediction errors in the ingroup compared
469 to the outgroup condition (**Fig. 3B**) is related to neural activation of the DLPFC/IFG. Using a
470 second-level regression analysis, we regressed the behavioral contrast between ingroup and
471 outgroup action learning weights κ against the neural contrast between the observation of
472 ingroup and outgroup choices, i.e., the time when observational action prediction error can be
473 computed. The results revealed only one significant whole brain corrected result, in the left IFG
474 (MNI_{xyz}: -34, 0, 28, Z_{stats}= 4.07, $P_{FWE\ whole\text{-}brain\ corrected}$ = 0.041) (**Figure 6A and B**). Thus, left IFG
475 activity reflected the impact of action prediction errors on behavior, which was reduced when
476 participants observed outgroup actions compared to when they observed ingroup actions. This
477 activity localized in the posterior part of the IFG region that was identified as the key region of
478 the action prediction error learning network.
479

480 **Discussion**

In this study, we investigated whether observational learning is shaped by the important social factor group membership. We report novel evidence that participants learn similarly well from observing ingroup and outgroup outcomes, but learn less well from observing outgroup actions. The observed deficit in learning from outgroup actions provides a plausible source of individuals' difficulties to learn from outgroup members, a phenomenon that has been described in previous studies (Buttelmann et al., 2013; Golkar et al., 2015; Golkar and Olsson, 2017), but so far not explained. Our neural results not only converge with the findings of previous observational learning studies in lateral prefrontal cortex (Burke et al., 2010), but also showed that IFG differentially encodes learning from observing ingroup vs outgroup actions.

In more detail, our behavioral findings revealed that participants made fewer correct choices after observing an outgroup demonstrator, compared to an ingroup demonstrator (Figure 2A) or to a computer demonstrator (Figure 2C). These outgroup deficits in observational learning occurred although participants observed comparable choice behavior in the ingroup and the outgroup condition and were able to predict ingroup and outgroup choices equally well. The finding of reduced learning from observing an outgroup compared to an ingroup individual is in line with previous behavioral evidence (Buttelmann et al., 2013; Golkar et al., 2015).

Extending these previous studies, we used computational learning models to specify the source of the outgroup deficit in observational learning. Given that observational learning is based on learning from observed outcomes (Burke et al., 2010; Suzuki et al., 2012; Kumaran et al., 2015) and observed actions (Burke et al., 2010; Suzuki et al., 2012), we hypothesized that outgroup deficits in observational learning might occur because individuals rely more on observed ingroup compared to outgroup outcomes, reflected by a stronger weight for ingroup compared to outgroup outcome prediction errors. Alternatively, we assumed that outgroup

deficits in observational learning might arise because participants rely more on observed ingroup compared to outgroup actions, reflected by a stronger weight for ingroup compared to outgroup action prediction errors. Our computational modeling results showed that the behavioral outgroup deficit in observational learning were best explained by a model in which participants put less weight on action prediction errors elicited by an outgroup individual than on action prediction errors elicited by an ingroup individual (Figure 3) and put similar weight on outcome prediction errors from the two groups.

The computational modeling results converge with the behavioral findings that showed a clear ingroup vs outgroup difference in the condition in which participants could only learn from actions, i.e., the action_only condition, and a marginal main effect of group ($F(1,28) = 3.32$, $p = 0.079$) in the condition in which participants could also learn from outcomes, i.e., the action_outcome condition. Presumably, the group effect in the action_outcome condition did not reach significance because the behavioral results reflect a mix of outcome-related, i.e., unbiased, and action-related, i.e., biased, learning. As computational models establish a relation between components of the phenomenon being modeled and the components of the model (Stafford, 2009), they can be more sensitive to the latent processes that might drive modulations in behavior than statistical analyses comparing behavioral outcomes alone (Stafford et al., 2020). In line with this notion, using computational modeling allowed us to disentangle the mixture of outcome- and action-based learning, and to specify the effect of group membership on the different subcomponents of observational learning.

Our neural results revealed that action- and outcome prediction errors elicited by observing the ingroup and the outgroup demonstrator are processed by similar neural circuitries that replicate previous findings. In more detail, learning from outcome prediction errors was

associated with activation in the DMPFC, the insula, the caudate and the midbrain, i.e., regions that have been implicated in outcome-related learning in previous neuroscience studies (Liu et al., 2011; Sescousse et al., 2013). Conversely, learning from action prediction errors was linked to neural responses in the anterior portion of the inferior frontal gyrus (IFG) and the parietal cortex, again in line with previous evidence (Burke et al., 2010; Suzuki et al., 2012). Key regions of the outcome-learning and action-learning networks that we obtained in the current studies showed considerable overlap with the respective neural circuitries observed in an independent previous study (Burke et al., 2010) that investigated observational learning independent of group membership (Figure 4).

Interestingly, although learning from ingroup and outgroup prediction errors activated similar neural networks, participants put stronger weight on the use of ingroup than outgroup action prediction error when they made decisions for themselves, which was reflected by stronger activation in the IFG (Figure 6). The IFG is involved in action-observation and imitation processes (Caspers et al., 2010) and forms part of the mirror neuron system (Molenberghs et al., 2012). Moreover, there is evidence that the activity of this area is modulated by group membership. For example, greater IFG activity was found when participants evaluated an ingroup member based on detailed personal information as compared to an outgroup member (Freeman et al., 2010). Another recent neuroimaging study revealed stronger activation in a mirror neuron network including left IFG when participants observed facial emotions of ingroup individuals compared to outgroup individuals (Krautheim et al., 2019). In line with this previous evidence, our results show that the processing of perceived actions in the IFG is modulated by group membership of the demonstrator during observational learning. Extending these previous findings, our results indicate that the IFG activity selects action-related information based on

social information (here group membership), and thus forms a plausible neural basis for biases in observational learning. However, the putative link with the mirror system will need to be tested formally.

It is worth noting the limitations of our study. First, we used exclusively political attitude to manipulate group membership. Given that previous research described outgroup learning deficits with group membership based on language or race (Buttelmann et al., 2013; Golkar et al., 2015), it is unlikely that outgroup learning deficits are limited to the political domain. Moreover, it is well-established that differences in political attitude foster social categorization, i.e., the formation of social ingroups and outgroups (Caruso et al., 2009; Rand et al., 2009; Young et al., 2014). There is even evidence that differences in political attitude can override social categorization based on race (Losin and Woo, 2015). In line with this previous evidence, the group induction based on political attitude in our study resulted in a salient group membership manipulation (Figure 1B), a conclusion further supported by our findings of significant behavioral and neural differences between the ingroup and the outgroup condition. That said, future research may want to investigate observational learning with a different group manipulation. Second, we studied left-wing participants only. Although targeting only one group (e.g. white participants in a study on race) to investigate ingroup-outgroup behavior is common in the literature (Golkar et al., 2015; Hein et al., 2016), future research may also want to study right-wing individuals or other political groups to generalize our findings. Third, we used a relatively small sample size and recent studies (Bossier et al., 2020; Marek et al., 2020) recommend larger sample sizes to ensure replication of fMRI findings than current practice (Yeung, 2018). It is therefore noteworthy that we replicate previous findings (Burke et al., 2010) on group-independent observational learning. Still, particularly our correlation findings should

be re-assessed with a larger sample in the future. Fourth, we did not specify the outgroup attributes that drive or shape the group differences in observational learning mechanisms revealed in our study. Future research may want to investigate whether the observed difference in observational learning arose because participants are less likely to trust outgroup actions without disambiguating feedback about the correctness of the outgroup demonstrator's choice (i.e., the outgroup outcome). Another factor that might play a role is the extent to which participants dislike the outgroup. In line with the findings of other studies (Golkar et al., 2015; Hein et al., 2016) it is conceivable that the individual impressions and/or emotions towards the respective outgroup might modulate the outgroup-related observational learning deficits observed in our study, an assumption that should be investigated in future studies.

In conclusion, the current results reveal that outgroup deficits in observational learning mainly reflect decreased learning from observed actions. Our findings suggest that the IFG differentially weighs ingroup and outgroup action prediction errors and provide a neurocomputational mechanism for outgroup deficits in observational action learning.

Data and code availability

The behavioral data is available online, https://osf.io/savw4/. Also, the neuroimaging results in this study can be found in https://neurovault.org/collections/VUOYOUFT/. Code to implement computational model is available from the corresponding author upon reasonable request.

References

Bossier H et al. (2020) The empirical replicability of task-based fMRI as a function of sample size. Neuroimage 212:1–12.

Burke CJ, Tobler PN, Baddeley M, Schultz W (2010) Neural mechanisms of observational learning. Proc Natl Acad Sci 107:14431–14436 Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1003111107.

597 Buttelmann D, Zmyj N, Daum M, Carpenter M (2013) Selective Imitation of In-Group Over
598 Out-Group Members in 14-Month-Old Infants. Child Dev 84:422–428.
599 Caruso EM, Mead NL, Balcetis E (2009) Political partisanship influences perception of biracial
600 candidates' skin tone. Proc Natl Acad Sci U S A 106:20168–20173.
601 Caspers S, Zilles K, Laird AR, Eickhoff SB (2010) ALE meta-analysis of action observation and
602 imitation in the human brain. Neuroimage 50:1148–1167 Available at:
603 http://dx.doi.org/10.1016/j.neuroimage.2009.12.112.
604 Charpentier CJ, Iigaya K, O'Doherty JP (2020) A Neuro-computational Account of Arbitration
605 between Choice Imitation and Goal Emulation during Human Observational Learning.
606 Neuron:1–13 Available at: https://doi.org/10.1016/j.neuron.2020.02.028.
607 Freeman JB, Schiller D, Rule NO, Ambady N (2010) The neural origins of superficial and
608 individuated judgments about ingroup and outgroup members. Hum Brain Mapp 31:150–
609 159.
610 Golkar A, Castro V, Olsson A (2015) Social learning of fear and safety is determined by the
611 demonstrator's racial group. Biol Lett 11:20140817 Available at:
612 http://rsbl.royalsocietypublishing.org/content/11/1/20140817.
613 Golkar A, Olsson A (2017) The interplay of social group biases in social threat learning. Sci Rep
614 7:1–5.
615 Hein G, Engelmann JB, Vollberg MC, Tobler PN (2016) How learning shapes the empathic
616 brain. Proc Natl Acad Sci 113:80–85 Available at:
617 http://www.pnas.org/lookup/doi/10.1073/pnas.1514539112.
618 Howard LH, Henderson AME, Carrazza C, Woodward AL (2015) Infants' and Young
619 Children's Imitation of Linguistic In-Group and Out-Group Informants. Child Dev 86:259–
620 275.
621 IBM Corp. (2015) IBM SPSS Statistics for Windows.
622 Krautheim JT, Dannlowski U, Steines M, Neziroğlu G, Acosta H, Sommer J, Straube B, Kircher
623 T (2019) Intergroup empathy: Enhanced neural resonance for ingroup facial emotion in a
624 shared neural production-perception network. Neuroimage 194:182–190.
625 Kumaran D, Warren DE, Tranel D (2015) Damage to the Ventromedial Prefrontal Cortex
626 Impairs Learning from Observed Outcomes. Cereb Cortex 25:4504–4518 Available at:
627 https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/bhv080.
628 Liu X, Hairston J, Schrier M, Fan J (2011) Common and distinct networks underlying reward
629 valence and processing stages: A meta-analysis of functional neuroimaging studies.
630 Neurosci Biobehav Rev 35:1219–1236 Available at:
631 http://dx.doi.org/10.1016/j.neubiorev.2010.12.012.
632 Losin EAR, Iacoboni M, Martin A, Cross KA, Dapretto M (2012) Race modulates neural activity
633 during imitation. Neuroimage 59:3594–3603.
634 Losin EAR, Woo C (2015) Brain and psychological mediators of imitation : sociocultural versus
635 physical traits. Cult Brain:93–111 Available at: http://dx.doi.org/10.1007/s40167-015-0029-
636 9.
637 Marek S et al. (2020) Towards Reproducible Brain-Wide Association Studies. bioRxiv 11:15–18
638 Available at: https://doi.org/10.1101/2020.08.21.257758.
639 Molenberghs P, Cunnington R, Mattingley JB (2012) Brain regions with mirror properties: A
640 meta-analysis of 125 human fMRI studies. Neurosci Biobehav Rev 36:341–349 Available at:
641 http://dx.doi.org/10.1016/j.neubiorev.2011.07.004.

642 R Core Team (2013) R: A language and environment for statistical computing. Available at:
643 http://www.r-project.org/.
644 Rand DG, Pfeiffer T, Dreber A, Sheketoff RW, Wernerfelt NC, Benkler Y (2009) Dynamic
645 remodeling of in-group bias during the 2008 residential election. Proc Natl Acad Sci U S A
646 106:6187–6191.
647 Sescousse G, Caldú X, Segura B, Dreher JC (2013) Processing of primary and secondary
648 rewards: A quantitative meta-analysis and review of human functional neuroimaging studies.
649 Neurosci Biobehav Rev 37:681–696 Available at:
650 http://dx.doi.org/10.1016/j.neubiorev.2013.02.002.
651 Suzuki S, Harasawa N, Ueno K, Gardner JL, Ichinohe N, Haruno M, Cheng K, Nakahara H
652 (2012) Learning to Simulate Others' Decisions. Neuron 74:1125–1137 Available at:
653 http://dx.doi.org/10.1016/j.neuron.2012.04.030.
654 Wunderlich K, Smittenaar P, Dolan RJ (2012) Dopamine Enhances Model-Based over Model-
655 Free Choice Behavior. Neuron 75:418–424.
656 Yeung AWK (2018) An updated survey on statistical thresholding and sample size of fMRI
657 studies. Front Hum Neurosci 12:1–7.
658 Young AI, Ratner KG, Fazio RH (2014) Political Attitudes Bias the Mental Representation of a
659 Presidential Candidate's Face. Psychol Sci 25:503–510.
660

661

662 **Table 1.** Summary of brain regions correlating with -sOPE as a parametric modulator in both
663 ingroup and outgroup conditions, whole brain cluster-level FWE- corrected p <0.05.

Brain region	Coordinates			T value	Z value	Voxels in cluster
	X	Y	Z			
DMPFC	-2	8	54	5.06	4.23	985
Caudate (L)	-16	10	6	4.30	3.73	283
Insula (L)	-42	16	4	4.73	4.02	176
Insula (R)	38	6	-8	4.62	3.95	263
Midbrain	-10	-36	-34	4.61	3.94	262
Precuneus	-26	-54	46	4.61	3.94	247

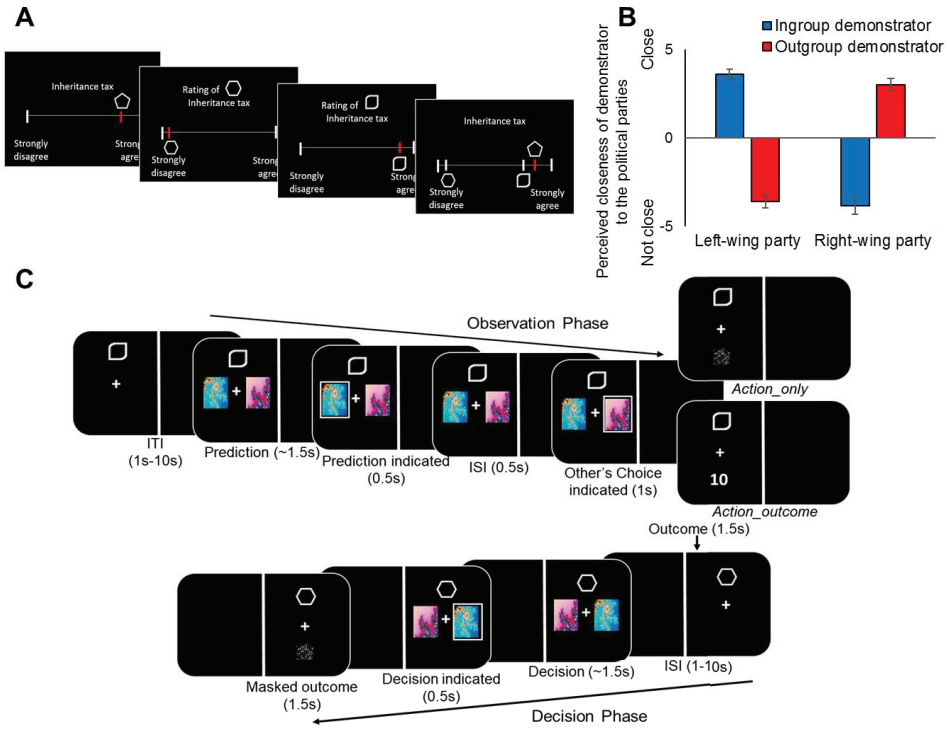
664
665
666 **Table 2.** Summary of brain regions correlating with APE as a parametric modulator in both
667 ingroup and outgroup conditions, whole brain cluster-level FWE- corrected p <0.05.

668

Brain region	Coordinates			T value	Z value	Voxels in cluster
	X	Y	Z			
Inferior Frontal gyrus (L)	-46	8	34	5.12	4.27	951
Inferior Frontal gyrus (R)	42	6	34	4.94	4.16	799
Precuneus/Inferior parietal lobe (R)	34	-68	40	4.88	4.11	467
Inferior parietal lobe (L)	-38	-48	44	4.79	4.06	654
Cerebellum /occipital lobe (R)	30	-62	0	5.70	4.61	1794
Cerebellum /occipital lobe (L)	-34	-74	-20	5.47	4.47	1463

669
670
671
672

673 **Figure captions**



674

675 **Figure. 1. Group induction, manipulation check, and observational learning task. A)** Prior
676 to scanning, participants rated current political issues in Switzerland and viewed ratings of two
677 different persons (future demonstrators) regarding the same issues. **B**) Participants' ratings of the
678 ingroup and outgroup demonstrator's closeness towards a left-wing and right-wing party, based
679 on how the demonstrators rated political initiatives (as in the example shown in Figure 1A). The
680 participants rated the ingroup demonstrator as similar to themselves, i.e., close to a left-wing
681 party and the outgroup demonstrator as dissimilar to themselves, i.e., close to a right-wing party.
682 Blue = perceived closeness of the ingroup demonstrator towards a left-wing (left panel) or right-
683 wing (right panel) party; red = perceived closeness of the outgroup demonstrator towards a left-
684 wing (left panel) or right-wing (right panel) party. Error bars indicate standard errors of the
685 mean. **C**) In the main part of the study, participants observed the ingroup or outgroup
686 demonstrator choosing between two fractal images. First, the demonstrator was indicated by one
687 of two abstract symbols (counterbalanced across participants). Then, participants had 1.5 s to

predict which option the demonstrator would choose. After response registration (0.5s) and a short interval (0.5s), the action of the demonstrator was shown for 1s. Depending on the condition, participants observed the outcome of the demonstrator (action_outcome condition) or a pixel-matched scrambled image (action_only condition) for 1.5 s. Next, the demonstrator's symbol was replaced by the participant's symbol, and the participants chose between the same options.

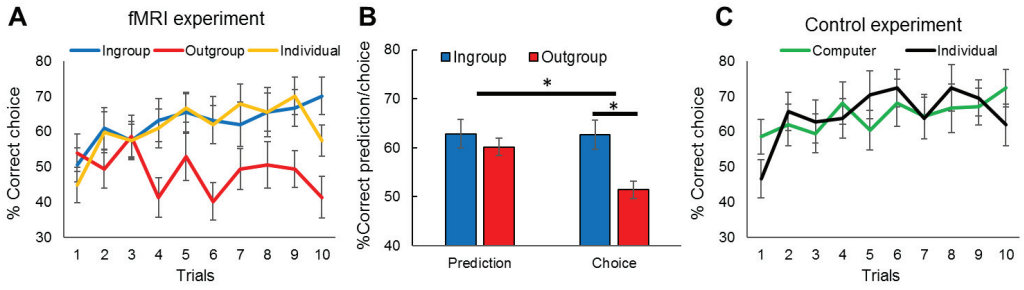
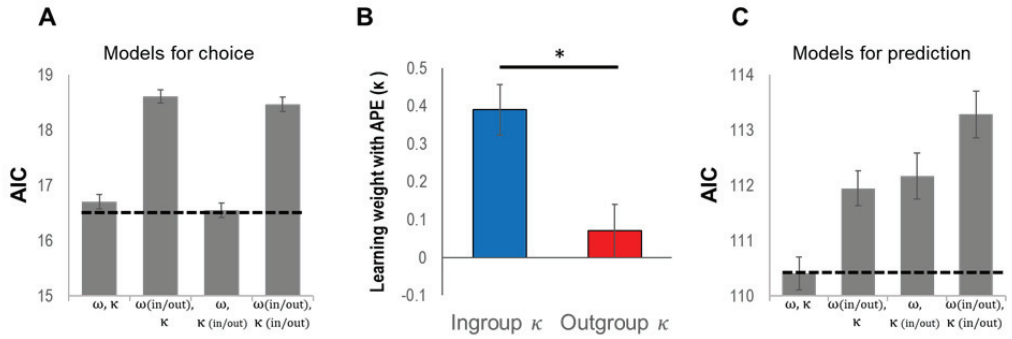


Figure 2. Behavioral results. A) Trial-wise percentage of correct choices in the three experimental conditions (ingroup, outgroup and individual learning) of the fMRI study **B)** Average percentage of correct choices and predictions. Group membership of the demonstrator affected correct choice but not prediction. **C)** Trial-wise percentage of correct choices in the two conditions (computer and individual learning) of the control experiment. Given that there were no significant differences between the action_only and the action_outcome conditions, the results are pooled over these two conditions in (A) and (C). Error bars indicate standard errors of the mean.



710

711 **Figure 3. Computational modeling results**. **A)** Model comparison based on the Akaike
712 information criterion (AIC) favored the model with a common outcome learning weight ϖ and
713 separate action learning weights κ for ingroup and outgroup demonstrators when participants
714 made decision for themselves. **B)** Ingroup vs outgroup difference in action learning weight (κ).
715 The weight given to ingroup action prediction errors (κ) was larger than the one given to
716 outgroup action prediction errors in behavior. Error bars indicate standard errors of the mean. **C)**
717 Model comparison for predictions of demonstrators' choices. AIC favored the model with a
718 common outcome learning weight ϖ and common action learning weights κ for ingroup and
719 outgroup demonstrators when participants predicted the decisions of the demonstrators.

720

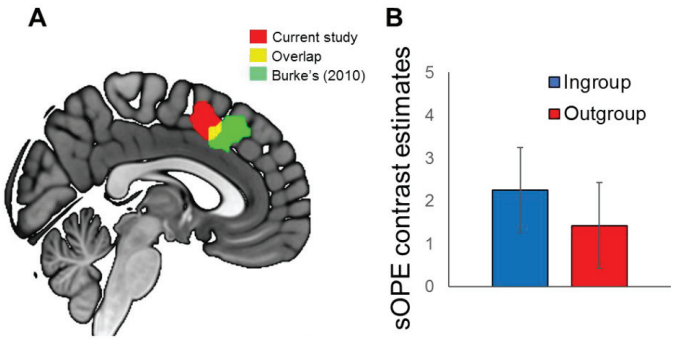
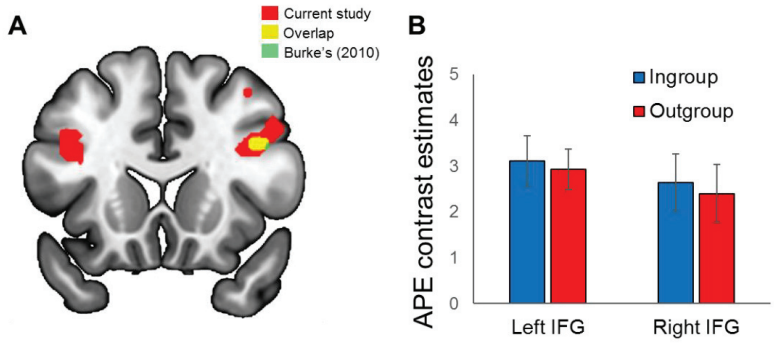


Figure 4. Outcome prediction error coding irrespective of group. A) Activation in dorsomedial prefrontal cortex (DMPFC) correlating with inverse simulated outcome prediction errors (sOPE) in both ingroup and outgroup conditions (red). This region overlapped (yellow) with the DMPFC region that correlated with inverse simulated outcome prediction errors in an independent previous study (green; unpublished data from Burke et al., 2010). For illustration purposes, results are displayed at p uncorrected < 0.001 (see **Table 1** for details and whole brain results). **B)** Bar plot of the DMPFC region shown in red (A), illustrating activity correlating with inverse simulated outcome prediction error for both ingroup and outgroup demonstrators. Error bars indicate standard errors of the mean.



737

738 **Figure 5. Action prediction error coding irrespective of group. A)** Prefrontal regions in left
739 and right IFG correlated with action prediction error (APE) as a parametric modulator regardless
740 of group (red). The region in right IFG overlapped with the DLPFC regions that correlated with
741 action prediction errors in a previous independent study (Burke et al., 2010; yellow / green). For
742 illustration purposes, results are displayed at p uncorrected < 0.001 (see **Table 2** for details and
743 whole brain results). **B)** Bar plots illustrating the relationship between the activity in left and
744 right IFG regions shown in (A) and ingroup and outgroup action prediction errors. Error bars
745 indicate standard errors of the mean.

746

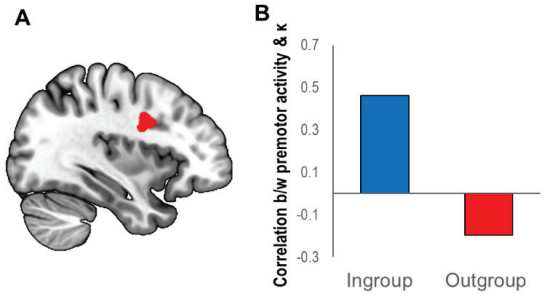


Figure 6. Group differences in neural activity correlating with behavioral effects. **A)** The ingroup vs. outgroup difference in κ correlated with ingroup vs. outgroup differences in inferior frontal gyrus (IFG, family wise cluster-level whole brain correction, $p = 0.041$, with an uncorrected voxel-level (i.e., cluster-inducing) threshold of $p < 0.001$). For illustration purposes, results are displayed at p uncorrected < 0.001. **B)** Correlation between individual IFG activity (extracted from the cluster) and κ.